APPENDIX 2.3: RULES OF PROBABILITY

The frequentist notion of probability is quite simple and intuitive. Here, we'll describe some rules that govern how probabilities are combined. Not all of these rules will be relevant to the rest of this book. However, describing these will help to make sure that we are using the concepts of probability correctly as we move on to more advanced topics.

We will begin with some notation. We can denote the probability of a flipped coin coming up heads as $p(heads) = .5$ and the probability of it coming up tails as $p(tails) = .5$. Or we can say that the probability of a rolled die coming up 1 is $p(1) = .1667$ and the probability of it coming up 3 is $p(3) = .1667$. However, we want to think about the general case of outcomes and events, not just those associated with coin flips or die rolls. Therefore, we will use letters to define arbitrary events. For example, we can use *A*, *B*, and *C* to denote three different events, no matter what variable we might be considering.

The OR Rule for Mutually Exclusive Events: $p(A \text{ or } B) = p(A) + p(B)$

A critical concept for us is the probability of *A* or *B* occurring. We've seen this question before, but now we can provide a bit more detail about how this is computed and what assumptions must be true for our calculation to be valid.

Events are *mutually exclusive* if they cannot co-occur. For example, a flipped coin can come up heads or tails, but not both. Therefore, the possible outcomes of a coin flip are mutually exclusive. Similarly, a rolled die can be one, and only one, of the following: 1, 2, 3, 4, 5, or 6. Therefore, these are mutually exclusive events. When we draw cards from a deck, the four suits are mutually exclusive. A drawn card can be a heart, but it can't simultaneously be a spade.

When events *A* and *B* are mutually exclusive, the probability of *A* **or** *B* occurring is the sum of their separate probabilities:

$$p(A \text{ or } B) = p(A) + p(B). \qquad (2.A3.1)$$

For example, if *A* and *B* are heads and tails, respectively, then the probability of a flipped coin being either a head (*A*) **or** a tail (*B*) is

$$p(A \text{ or } B) = p(A) + p(B) = .5 + .5 = 1.$$

If we consider the role of a die, and *A* and *B* are 4 and 6, respectively, then the probability of a rolled die coming up 4 **or** 6 is

$$p(A \text{ or } B) = p(A) + p(B) = \frac{1}{6} + \frac{1}{6} = .33.$$

Or if we consider the role of a die and *A*, *B*, and *C* are 1, 4, and 6, respectively, then the probability of a rolled die coming up 1 **or** 4 **or** 6 is

$$p(A \text{ or } B \text{ or } C) = p(A) + p(B) + p(C) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = .5.$$

The OR rule is the most important rule of probability for much of what follows in subsequent chapters.

The AND Rule for Independent Events: $p(A \text{ and } B) = p(A)p(B)$

Two events (or outcomes) are *independent* if the occurrence of one does not affect the probability that the other will occur. For example, if two coins are flipped, the outcomes are independent. In other words, if one coin comes up heads, it has no effect on whether the other coin will come up heads. Or if the same coin is flipped twice, coming up heads on the first flip has no effect on the probability of it coming up heads on the second flip. Each time a coin is flipped, the outcome is independent of the outcomes of all previous flips.

When events *A* and *B* are independent, the probability of *A* **and** *B* occurring is the *product* of their separate probabilities:

$$p(A \text{ and } B) = p(A)\,p(B). \qquad (2.A3.2)$$

For example, if *A* and *B* are heads and tails, respectively, then the probability of flipping a coin twice and getting a head (*A*) on the first flip **and** a head (*B*) on the second flip is

$$p(A \text{ and } B) = p(A)p(B) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right) = (.5)(.5) = .25.$$

Notice that each of the following events has the same probability of occurrence: (*head* and *tail*), (*head* and *head*), (*tail* and *tail*), and (*tail* and *head*). These are the four possible outcomes for two flips of a coin, and each has a probability of .25. The sum of these four probabilities is 1, because no other outcomes are possible. In this example, we've considered two successive flips of the same coin, but the result would be exactly the same if we considered flipping two coins simultaneously.

The OR Rule for Events That Are Not Mutually Exclusive: $p(A \text{ or } B) = p(A) + p(B) - p(A)p(B)$

Some events are not mutually exclusive. For example, a card drawn from a deck can be both a Heart and a

King. A student can be both female and in psychology. A person can be both anxious and depressed. When events are not mutually exclusive, the OR rule is modified as follows:

$$p(A \text{ or } B) = p(A) + p(B) - p(A)\,p(B). \qquad (2.A3.3)$$

Equation 2.A3.3 differs from equation 2.A3.1 only in the last term, $p(A)p(B)$, which denotes the probability of both A and B occurring.

Let's consider drawing a card from a 52-card deck that has four suits (Clubs, Spades, Hearts, and Diamonds) and 13 ranks (Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, and King). If event A is "drawing a red card" and event B is "drawing a King," then we can ask about the probability of A or B. These events are not mutually exclusive. If you draw a red card, it could be a King. Conversely, if you draw a King, it could be red. Figure 2.A3.1 shows a full deck of playing cards to help us think about this question. The bottom two rows show all the red cards; diamonds and hearts. These represent half the deck, so the probability of drawing a red card is $p(A) = .5$. The right column shows the four Kings. Because four of the 52 cards are Kings, the probability of drawing a King is $p(A) = 4/52 = 1/13 = .07692$. Because A and B are not mutually exclusive, we have to take into account the probability that a card is both red and a King. The probability of being red and being a King is

$$p(A)\,p(B) = \left(\frac{1}{2}\right)\left(\frac{1}{13}\right) = \frac{1}{26} = .03846.$$

Another way to say this is that red Kings compose 1/26th of the deck.

Equation 2.A3.3 tells us that we should do the following to calculate the probability of drawing a card that is red or a King:

$$\begin{aligned} p(A \text{ or } B) &= p(A) + p(B) - p(A)\,p(B) \\ &= \frac{1}{2} + \frac{1}{13} - \left(\frac{1}{2}\right)\left(\frac{1}{13}\right) = \frac{1}{2} + \frac{1}{13} - \left(\frac{1}{26}\right) \\ &= .5 + .07692 - .03846 = .53846. \end{aligned}$$

We can confirm that this is the correct answer by counting the number of cards that satisfy our two *constraints* of being red or being a King. There are 26 red cards, including the red Kings. When we add in the two black Kings, we now have 28 cards altogether. Therefore, the proportion of cards that satisfy conditions A or B is 28/52 = 7/13 = .53846. We can now see that subtracting the third term in equation 2.A3.3, $p(A)p(B)$, from the first two serves to prevent red Kings from being counted twice.

The AND Rule for Dependent Events: $p(A \text{ and } B) = p(A)p(B|A)$

Not all events are independent; some are *dependent*. To understand dependence, let's first think about *independent* events. Let's say we draw a card from a shuffled deck, put it back in, reshuffle, and then draw again. This is called *sampling with replacement.* What is the probability of drawing two aces in two successive draws when sampling with replacement? Well, there are two events (A = drawing an Ace on the first draw, B = drawing an Ace on the second draw). The probability of A is $p(A) = 1/13$ and the probability of B is $p(B) = 1/13$. Therefore, using the AND rule (from equation 2.A3.2), we find that the probability of A and B is $p(A \text{ and } B) = p(A)p(B) = 1/(13 * 13) = 1/169 = .00592$.

Now, let's change the example slightly and imagine drawing two cards without replacing the first one before the second one is drawn. This is called *sampling without replacement.* What is the probability now of drawing two aces? If an Ace had been drawn on the first draw, then the probability of an Ace on the second draw has changed. If an Ace was the first card drawn, then only 51 cards remain and only three of these are aces. Therefore, the probability of drawing an Ace on the second draw *depends* on whether an Ace was drawn on the first draw. Therefore, we can't use equation 2.A3.2. Rather, we use equation 2.A3.4 as follows:

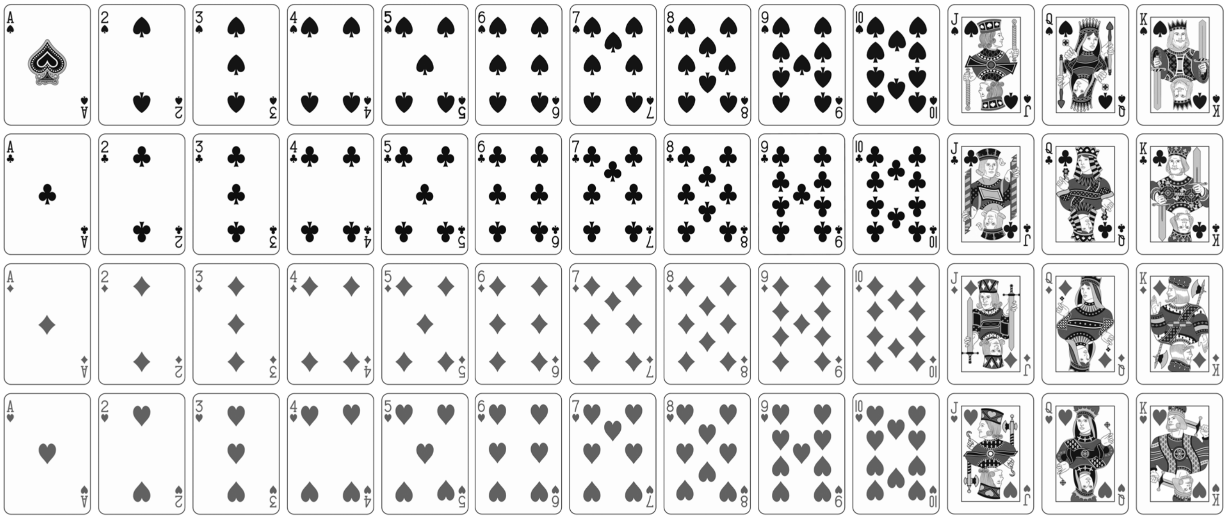
$$p(A \text{ and } B) = p(A)\,p(B \mid A). \qquad (2.A3.4)$$

The term $p(B|A)$ should be read as "the probability of event B occurring, given that event A has occurred." In our example, this means the probability of drawing an Ace on the second draw, given that an Ace was drawn on the first draw. We call $p(B|A)$ a *conditional probability.*[1]

Because there are four aces in the deck, the probability of the first card drawn being an Ace is $p(A) = 4/52 = 1/13 = .07692$. As we noted, if the first card drawn was an Ace, then there are only three aces in the remaining 51 cards. So, when the second card is drawn, the probability of drawing an Ace is only $p(B|A) = 3/51$

[1] Please note, we will return to the important issue of conditional probabilities in Chapter 7, where we discuss significance tests. If you hear that a result is statistically significant, this means someone has conducted a significance test. You may be surprised to learn that psychologists are often harshly criticized for misinterpreting the results of significances tests. Many of these misinterpretations arise from not understanding the concept of conditional probability. Therefore, conditional probability is not a minor concept. It is hugely important for the correct interpretation of significance tests. See you in Chapter 7.

FIGURE 2.A3.1 ■ A Deck of Playing Cards



There are four suits (Spades, Clubs, Diamonds, and Hearts) and 13 ranks (Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, and King).

© iStock.com/imannaggia

= 1/17 = .05882. If we now work through equation 2.A3.4, we will find that

$$p(A \text{ and } B) = p(A)p(B \mid A) = \left(\frac{1}{13}\right)\left(\frac{3}{51}\right)$$
$$= \frac{3}{663} = .00452.$$

So, the probability of drawing two aces is greater if we draw with replacement than if we draw without replacement. Another way to say this is that the probability of drawing two aces is greater when the draws are independent versus dependent.

LEARNING CHECK 1

1. What is the probability that a card drawn from a 52-card deck will be an 8 or a 9?
2. What is the probability that in two independent draws from a 52-card deck, the first card will be an 8 and the second card will be a 9?
3. What is the probability that a card drawn from a 52-card deck will be an 8 or red?
4. What is the probability that in two successive draws from a 52-card deck, the first card will be an 8 and the second will be a 9 when sampling is without replacement?

Answers

1. $p = p(8) + p(9) = 4/52 + 4/52 = 8/52 = 2/13 = .1538.$
2. $p = p(8) * p(9) = 4/52 * 4/52 = 1/13 * 1/13 = 1/169 = .0059.$
3. $p = p(8) + p(\text{red}) - p(8)p(\text{red}) = 1/13 - 1/2 - (1/2 * 1/13) = .5385.$
4. $p = p(8) * p(9|8) = 1/13 * 4/51 = .0769 * .0784 = .006.$

APPENDIX 2.4: PROBABILITY DENSITY FUNCTIONS

Functions

You probably encountered *functions* in high school mathematics. If not, then you almost certainly recognize this: $y = x^2$. This is the *square function*. Functions are like black boxes. You put a number in, and you get a number out. For this reason, it's common to express functions like this: $y = f(x)$. The f means function, x is the input, and y is the output. Something goes on inside the black box called f, and a number pops out, which we call y. In the case of the square function, you put in some number x, and you get out the square of the number, which we call y. The defining feature of a function is that there is a single y value for every possible x value. Therefore, y is said to be a function of x. Probability density functions are functions for this reason; there is a single y value for each x value, as shown in Figure 2.4. But what *is* the y value in Figure 2.4?

Density

The term *density* should be familiar. When we talk about population density, for example, we mean the number of people per square mile or square kilometer. Population density is greater in cities than in rural areas. Density usually refers to the number of things (people, trees, worms, neurons) per unit measure (square mile, acre, cubic foot, cubic millimeter). In a grouped frequency table (e.g., Table 2.8), we can think of the number of scores per interval as density. The more scores per interval, the greater the density. So, the raw frequency counts tell us something about the density of scores in an interval.

The notion of density is more abstract for mathematicians and statisticians. It differs from the usual notion of density in that it is defined at a point rather than for some width, area, or volume. How can density be defined at a point? Let's start by thinking about a traffic jam that stretches for 5 miles, or 8.05 kilometers. Cars are packed bumper to bumper, so the density of cars is the same at each point along the highway. If you count the number of cars in a 1-kilometer stretch (interval), you might find that there are 400 cars in this interval. So, the density is 400 per kilometer. If you count the number of cars in a half-kilometer interval, you would find 200 per half kilometer. Now, 400 per kilometer is the same density as 200 per half kilometer, and it is also the same as 100 per quarter kilometer. All of these measures of density can be put on the same scale by dividing the number of cars in an interval by the interval width. The interval widths for this example are 1 kilometer, .5 kilometers, and .25 kilometers. So, if we divide the counts (400, 200, and 100) by the corresponding interval widths, we obtain $400/1 = 400$, $200/.5 = 400$, and $100/.25 = 400$. In this way, density can be computed independently of interval width. So, how does this relate to specifying density at a point?

We will now return to the distribution of heights that we discussed in Chapter 2. Figure 2.A4.1 shows histograms of 1,000,000 heights drawn from a known distribution. The widths of the intervals decrease from 5.33 inches (Figure 2.A4.1a) to .67 inches (Figure 2.A4.1f). As interval width decreases, fewer scores fall in each interval. Therefore, the heights of the histogram bars decrease as interval width decreases.

In our traffic jam example, we noted that density involves dividing the number or proportion of scores in each interval by the interval width. This has been done in Figure 2.A4.2, in which the bar heights ($p = n/N$) from Figure 2.A4.1 are divided by the interval width (p/width) to yield density. As interval width decreases, the tops of the histogram bars become indistinguishable from the solid line, which represents the probability density function of the distribution from which the scores were drawn.

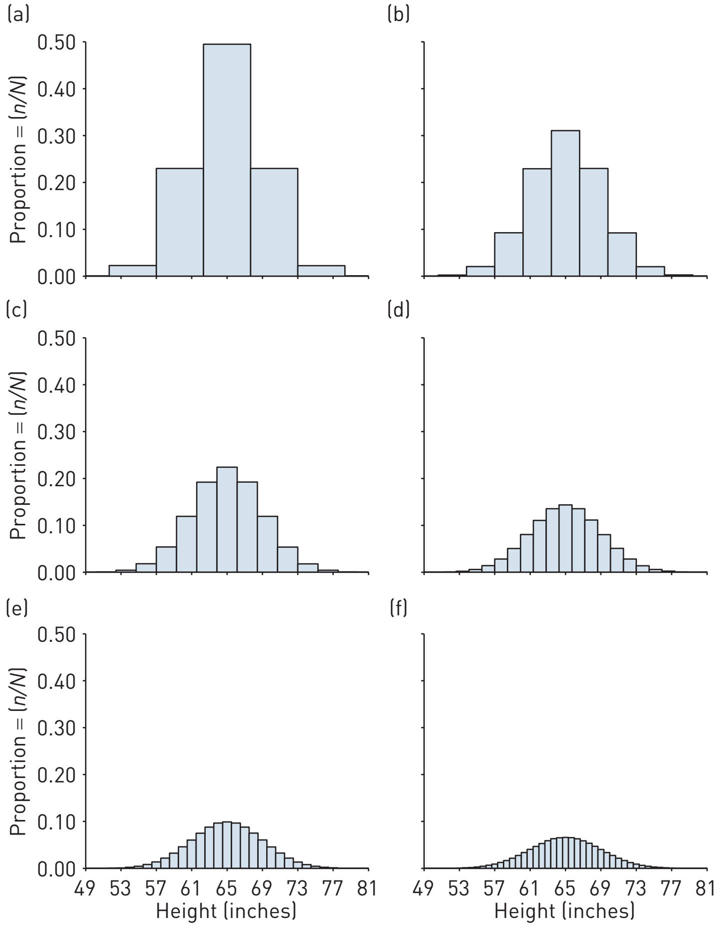
Let's now think of a theoretical population with an infinite number of scores, rather than *just* 1,000,000. As interval width becomes smaller and smaller, two things happen. First, density converges to a single unambiguous value. (To see this, think about the histogram bar centered on 65 in Figures 2.A4.2a through 2.A4.2f.) Second, in the limit, the width becomes zero. This means that (i) density can be defined at a point and (ii) the probability of any specific score actually occurring is 0. The result is the continuous line (function) that defines a y value (density) for each x value. We call this a probability density function.

It might seem like a bit of a paradox that as interval width decreases, the density of scores in a small region of the distribution approaches a constant value, whereas the proportion of scores in each interval approaches zero. This is something we just have to live with.

Probability Density

So far, we've seen the following things. The curve in Figure 2.4 is a function. The y values represent the abstract notion of density defined at a point. Density does not mean probability. So why is this called a *probability* density function? Let's see if we can answer this.

FIGURE 2.A4.1 ■ Histograms of 1,000,000 Heights



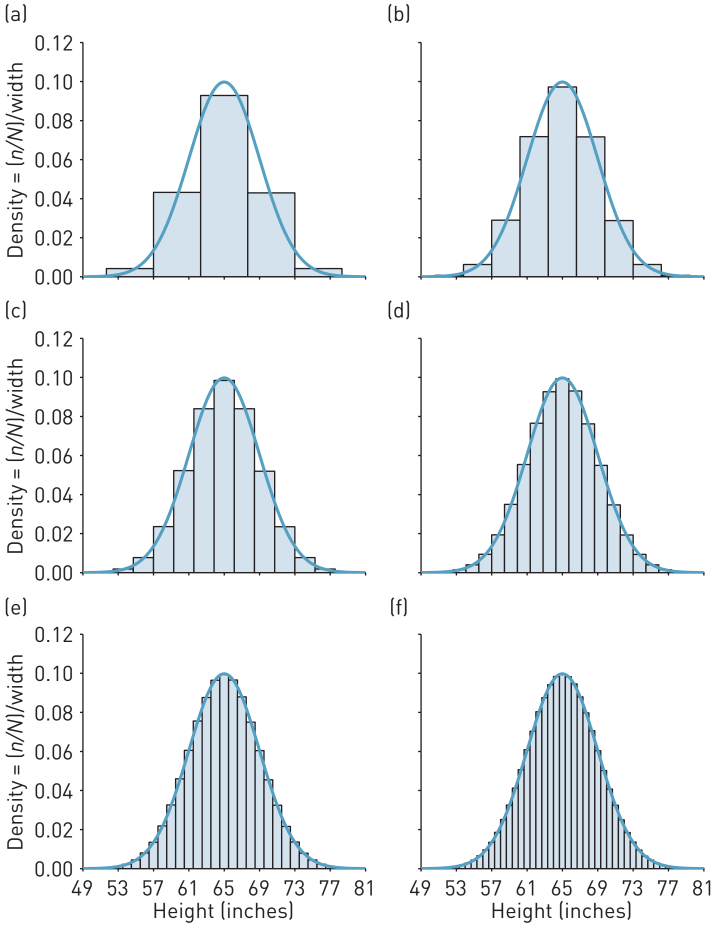
(a through f) Each panel shows a histogram of 1,000,000 heights. The interval widths range from 5.33 inches (a) to .67 inches (f). The y-axis represents the proportion of scores ($p = n/N$) in each interval. As interval width decreases, fewer scores fall in each interval.

In Chapter 2, we considered distributions defined by the categories of qualitative variables, discrete values of quantitative variables, and intervals of quantitative variables. Each of these categories or intervals was associated with a probability, and the sum of all these probabilities is 1. Something very similar is true of a probability density function. As interval widths get narrower, the number of intervals increases while the proportion of scores in each interval decreases. This means that no matter how narrow the intervals, the sum of the proportions in the intervals will be 1. So, here is another oddity for us. As the interval width approaches zero, the sum of the proportions associated with the intervals remains 1. At the same time, no matter how narrow the intervals are, some will contain more scores than others. This is another seeming paradox that we just have to live with.

If you have taken a calculus course, you will recognize that I've just described *integration*. Therefore, we can say that density functions are probability functions, because the area under the curve is 1. For this reason, the function in Figure 2.A4.2 (the curved line) is a probability function. If we compute the area under the curve between any two values of x, we obtain the probability that a randomly chosen score will fall in that interval.

And that's all I have to say about that.

FIGURE 2.A4.2 ■ An Illustration of Density



(a through f) Densities computed for 1,000,000 heights. The interval widths range from 5.33 inches (a) to .67 inches (f). The *y*-axis represents the proportion of scores ($p = n/N$) in each interval divided by the width of the interval *p*/width. The solid line is the mathematical density function associated with the distribution from which the scores were drawn. As the interval width approaches 0, the heights of the histogram bars increasingly resemble the continuous probability density function or *pdf*.